

<https://helda.helsinki.fi>

---

## Can We Assume the Gene Expression Profile as a Proxy for Signaling Network Activity?

Piran, Mehran

Multidisciplinary Digital Publishing Institute

2020-06-03

---

Piran, M.; Karbalaee, R.; Piran, M.; Aldahdooh, J.; Mirzaie, M.; Ansari-Pour, N.; Tang, J.; Jafari, M. Can We Assume the Gene Expression Profile as a Proxy for Signaling Network Activity? *Biomolecules* 2020, 10, 850.

---

<http://hdl.handle.net/10138/316176>

---

---

*Downloaded from Helda, University of Helsinki institutional repository.*

*This is an electronic reprint of the original article.*

*This reprint may differ from the original in pagination and typographic detail.*

*Please cite the original version.*

## Article

# Can We Assume the Gene Expression Profile as a Proxy for Signaling Network Activity?

Mehran Piran <sup>1</sup>, Reza Karbalaie <sup>2</sup>, Mehrdad Piran <sup>3</sup>, Jihad Aldahdooh <sup>4</sup>, Mehdi Mirzaie <sup>5</sup>, Naser Ansari-Pour <sup>6</sup>, Jing Tang <sup>4,\*</sup> and Mohieddin Jafari <sup>4,\*</sup>

<sup>1</sup> Bioinformatics and Computational Biology Research Center, Shiraz University of Medical Sciences, Shiraz P.O. Box 71336-54361, Iran; piranmehran@gmail.com

<sup>2</sup> Department of Biology, Temple University, Philadelphia, PA 19122, USA; reza.karbalaie@temple.edu

<sup>3</sup> Department of Tissue Engineering and Applied Cell Sciences, School of Advanced Technologies in Medicine, Shahid Beheshti University of Medical Sciences, Tehran 14177-55469, Iran; mehrdadp1989@gmail.com

<sup>4</sup> Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, 00270 Helsinki, Finland; jihad.alдахdooh@helsinki.fi

<sup>5</sup> Department of Applied Mathematics, Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran P.O. Box 14115-134, Iran; mirzaie@modares.ac.ir

<sup>6</sup> Nuffield Department of Medicine, Big Data Institute, University of Oxford, Oxford OX3 7LF, UK; naser.ansari-pour@bdi.ox.ac.uk

\* Correspondence: jing.tang@helsinki.fi (J.T.); mohieddin.jafari@helsinki.fi (M.J.)

Received: 12 May 2020; Accepted: 31 May 2020; Published: 3 June 2020

**Abstract:** Studying relationships among gene products by expression profile analysis is a common approach in systems biology. Many studies have generalized the outcomes to the different levels of central dogma information flow and assumed a correlation of transcript and protein expression levels. However, the relation between the various types of interaction (i.e., activation and inhibition) of gene products to their expression profiles has not been widely studied. In fact, looking for any perturbation according to differentially expressed genes is the common approach, while analyzing the effects of altered expression on the activity of signaling pathways is often ignored. In this study, we examine whether significant changes in gene expression necessarily lead to dysregulated signaling pathways. Using four commonly used and comprehensive databases, we extracted all relevant gene expression data and all relationships among directly linked gene pairs. We aimed to evaluate the ratio of coherency or sign consistency between the expression level as well as the causal relationships among the gene pairs. Through a comparison with random unconnected gene pairs, we illustrate that the signaling network is incoherent, and inconsistent with the recorded expression profile. Finally, we demonstrate that, to infer perturbed signaling pathways, we need to consider the type of relationships in addition to gene-product expression data, especially at the transcript level. We assert that identifying enriched biological processes via differentially expressed genes is limited when attempting to infer dysregulated pathways.

**Keywords:** gene expression; signaling network; network biology; transcriptomics; differentially expressed genes; causality analysis

## 1. Introduction















In network biology, defining causal relationships among nodes is crucial for static and dynamic analysis [1,2]. The most available high-throughput data for inferring molecular relationships are arguably whole-transcriptome expression profiles analyzed with statistical models [3]. The challenge is extrapolating causality in signaling and regulatory mechanisms from a significant correlation

between any given gene pair. Many spurious correlations among gene pairs lacking any direct causal relationship are possible, both stochastically and as a result of indirect effects [4]. So far, reverse engineering algorithms have been developed to tackle this challenge and to infer gene networks and regulatory interactions from expression profiles [5–7].

When considering signaling networks as a portrait of causal relationships among the molecular entities in biology, their leading players are proteins, whose activity is often regulated by post-translational modifications such as phosphorylation and acetylation [8,9]. Therefore, inference of signaling networks can be directly inferred from (Phospho) proteomic and protein–protein interaction data [10]. However, these kinds of data are expensive and laborious to acquire. Given the correlation between protein and gene expression, a common alternative approach is using gene expression to infer interactions between proteins. It is known that gene expression or transcriptome refers to “what appears to happen in a biological system”, while the signaling network explicates “what makes it happen and what has happened in a complex view of the system” [11]. Meanwhile, in transcriptomic studies, differentially expressed genes (DEGs) are identified based on the changes in mRNA levels between two or more groups of samples, and several affected biological pathways are recognized by node-centric enrichment analysis [12,13]. This, therefore, begs the question of whether gene expression profiles amplify the mechanism of signaling circuits, i.e., activatory/inhibitory relationships.

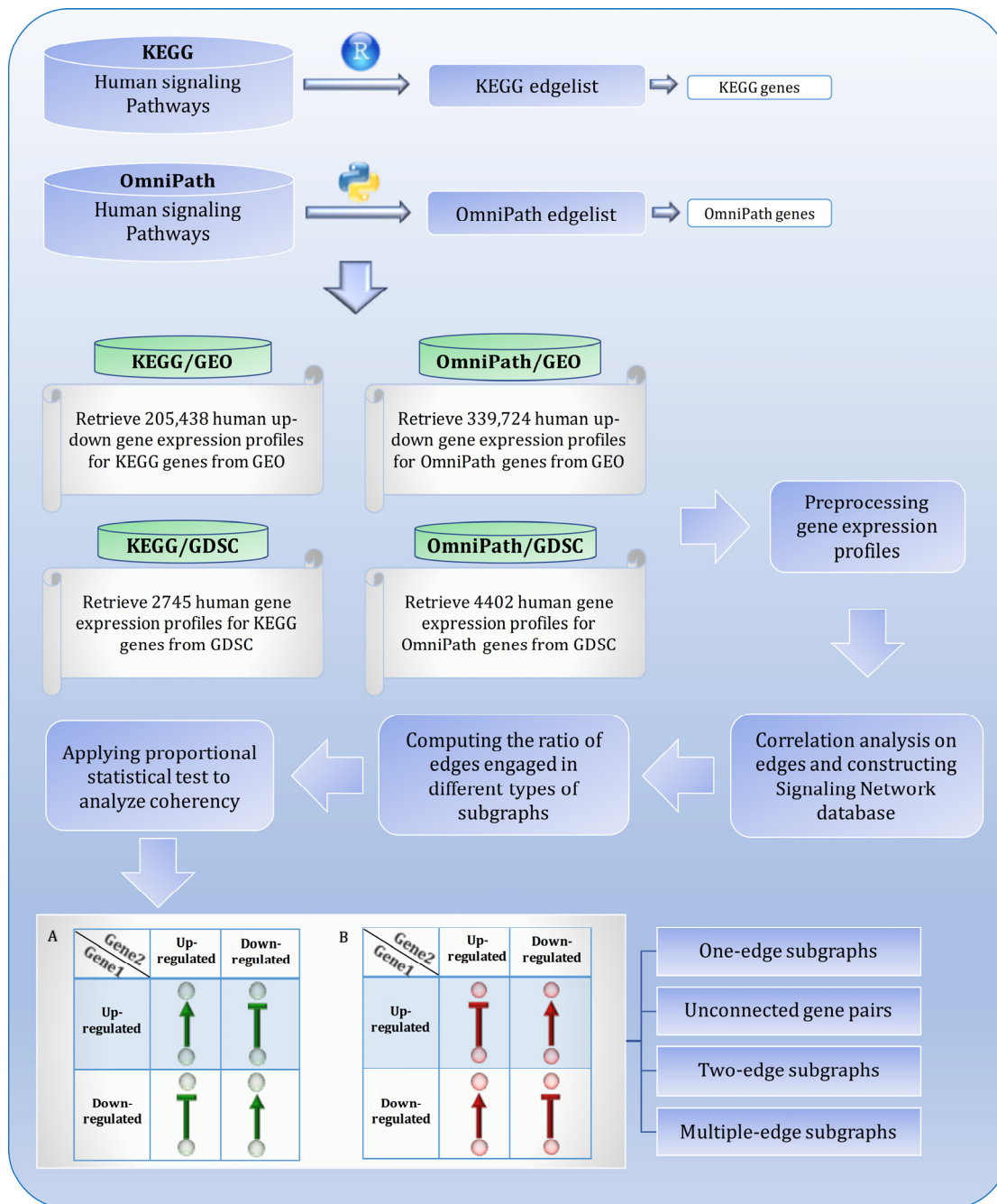
In this study, we aimed to examine the coherency between expression profiles and the signs of relationship in signaling networks for all possible gene pairs. To elaborate on what we mean by coherency, imagine a gene pair ( $G_I$ ,  $G_{II}$ ) where gene  $G_I$  activates gene  $G_{II}$ . If the expression profiles of both are correlated positively, we infer that expression data strengthen the activation type of this signaling relationship and are thus a coherent gene pair. In contrast, let gene  $G_I$  inhibit gene  $G_{II}$ . In this case, the coherent gene pairs are negatively correlated, and the inhibition type of this signaling relationship is strengthened in accordance with the expression levels of transcripts. If gene  $G_I$  activates gene  $G_{II}$  while there is a negative correlation between them or if gene  $G_I$  inhibits gene  $G_{II}$  while there is a positive correlation between them, this implies the incoherency between the gene pair relationships. In other words, incoherent gene pairs’ effects on the action of each other are not supported by their mRNA expression levels. In addition to these simple scenarios, we also considered more complicated subgraphs in a signaling network (See Table 1). Edges appear between gene pairs, indicating a relationship. Since all edges have a direction, the type of logical relationship is determined using the sign of edges. We then performed a correlation analysis between the connected gene pairs compared with unconnected gene pairs to infer coherency. To this end, we used expression datasets in the Gene Expression Omnibus (GEO) [14] and Genomics of Drug Sensitivity in Cancer (GDSC) databases [15] to extract the relevant gene expression profiles. Two literature-curated databases for signaling pathways, namely the Kyoto Encyclopedia of Genes and Genomes (KEGG) [16] and OmniPath [17] (which integrates literature-curated human signaling pathways from 34 resources) were used to extract the sign of relationships among directly linked gene pairs. Therefore, a coherency analysis was undertaken independently for all four combinations of databases in parallel (see Table 1). This analysis allowed us to explore the importance of having coherent gene pair relationships to infer the dysregulated signaling pathways.

**Table 1.** Details of different subgraphs present in all biological signaling networks. The dashed lines indicate multiple edges between nodes. The last two columns provide the number of each subgraph in the two signaling databases.

Simple Subgraphs					
Structures	Names	Abbrev.	KEGG	OmniPath	
	Unconnected Gene Pairs	UGP	—	—	
	Activation	Act	19,170	15,841	
	Inhibition	Inh	7320	5012	
Complex Subgraphs					
	Dual Negative Feedback Loop	DNFBL	37	279	
	Dual Positive Feedback Loop1	DPFBL1	186	912	
	Dual Positive Feedback Loop2	DPFBL2	14	173	
	Multiple Negative Feedback Loop1	MNFBL1	17,712	14,913	
	Multiple Positive Feedback Loop1	MPFBL1	3731	4104	
	Multiple Negative Feedback Loop2	MNFBL2	2417	1005	
	Multiple Positive Feedback Loop2	MPFBL2	3232	3279	
	Multiple Feed-Forward Loop1	MFFL1	12,869	6729	
	Multiple Feed-Forward Loop2	MFFL2	6618	4718	
	Multiple Negative Feed-Forward Loop1	MNFFL1	8918	9663	
	Multiple Negative Feed-Forward Loop2	MNFFL2	2925	842	

## 2. Materials and Methods

In this study, four independent analyses were performed based on two gene expression databases, i.e., GEO and GDSC, and two signaling pathway databases, i.e., KEGG and OmniPath, in parallel (Figure 1) [14–17]. The signaling pathway databases were thus independently used to reconstruct a whole signaling network, and the gene expression databases were separately used to apply a correlation analysis to each gene pair in the pathways to undertake and compare the distinct analyses and findings of KEGG/GEO, KEGG/GDSC, OmniPath/GEO, and OmniPath/GDSC. To briefly introduce the utilized gene expression databases, GEO is an international public repository of National Center for Biotechnology Information (NCBI) that archives microarray and next-generation sequencing expression data. The GDSC database is the largest public repository that archives information about drug sensitivity in cancer cells and biomarkers of drug response in these cells. In this work, gene expression profiles from GDSC cell lines and GEO studies were used to extract pairwise association between genes.



**Figure 1.** Visual overview of how information from different databases was integrated to analyze the coherency. An edge list was constructed from Kyoto Encyclopedia of Genes and Genomes (KEGG) and OmniPath databases. All the gene expression profiles for the edge list genes were then downloaded from Gene Expression Omnibus (GEO) and Genomics of Drug Sensitivity in Cancer (GDSC) databases. Next, data were preprocessed, and a suitable structure was created for correlation analysis among the gene pairs. By interpreting the information from correlation tests and the proportional tests, coherency analysis was implemented on different forms of subgraphs. There are a total of four coherent conditions in panel A and four incoherent conditions in panel B. For instance, in panel A, if gene1 is up-regulated and there is an activation between the gene pair, gene2 must be up-regulated. In panel B, if gene1 is up-regulated and there is an inhibitory relationship between the gene pair, gene2 is expected to be up-regulated.

### 2.1. Signaling Network Reconstruction

Here, we focused on human signaling pathways based on available datasets. All human-related signaling pathways were downloaded from the KEGG database. Using the KEGGgraph package [18], these pathways were imported into the R environment [19]. Edge information was extracted, and each graph was converted to an edge list. Next, all edges ( $n = 26,490$ ) were merged, and a directed signed signaling network was reconstructed (Supplementary File 1 (Section 1), and Supplementary File 4). Eligible edges (see Section 2.3) were then selected, and correlation analysis was undertaken on eligible gene pairs. The pypath python module [17] was also used to do the same and create an edge list based on the OmniPath database (see Supplementary File 4). This edge list ( $n = 20,853$ ) was also imported into the R environment for the downstream statistical analysis on the gene pairs.

### 2.2. Gene Expression Profiles Extraction

The standard GEO query format (GEO Profiles) was used to identify all up- and down-expressed genes or DEGs represented within the KEGG and/or OmniPath edge lists. For each study, a number of differentially expressed genes are identified based on the desired thresholds of log fold change and  $p$ -values indicating significance. As the up- and downregulated gene expression profiles are derived from different kinds of studies corresponding to different periods and different genomic technologies, we relied on the definition of up- and downregulated genes in the GEO database to define DEGs. Gene expression profiles available in GDSC were downloaded for both edge lists, followed by preprocessing and outlier detection. Finally, based on GEO and GDSCs, four expression matrices were created using the genes which make up of KEGG and OmniPath edge list (Supplementary File 1 (Sections 2 and 3) and Supplementary File 5). For more details about the total number of gene expression profiles, DEGs, and total number of samples that were downloaded from GEO and GDSC, see Figure 1 and Table 2.

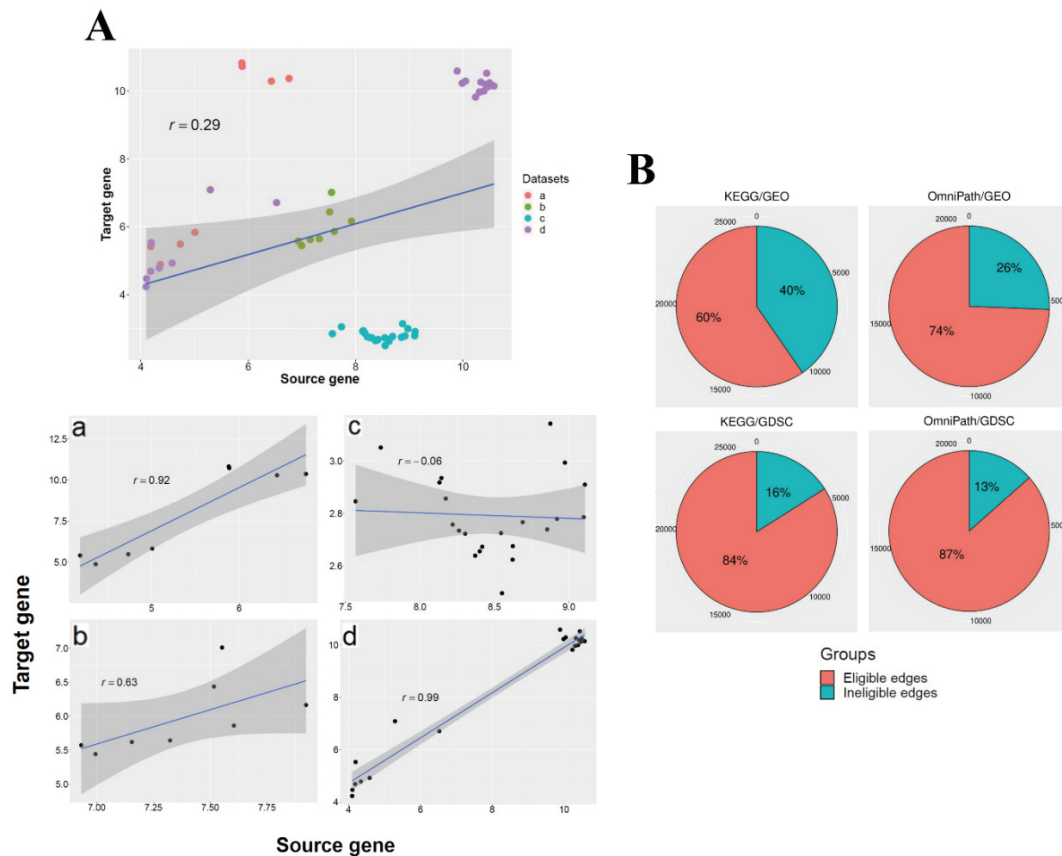
### 2.3. Mutual Association Analysis

In the next step, the correlation of the expression profiles of each gene pair was statistically tested. For computing correlation coefficients for any gene pair, including Pearson, Spearman's rank, and Kendall rank coefficients, we only considered gene pairs having more than two samples. These gene pairs were considered as eligible edges for downstream statistical analysis (Supplementary File 1 (Section 4)). Samples with expression data for the gene pairs may have come from different datasets with multiple organisms and tissues and therefore they were analyzed separately and independently. Figure 2A represents the effect of this preprocessing on an exemplary gene pair in our dataset. In this study, the gene expression profiles were considered dataset-specific, as mentioned, to avoid any inconsistency among the samples collected from diverse datasets, as sample heterogeneity can easily obscure pairwise relationships. An edge is therefore considered homogeneous if the correlation sign is consistent throughout. These homogeneous edges were used for correlation analysis. Then, according to the statistical significance and the sign of the correlation coefficient, the coherent and incoherent gene pairs were inferred (Supplementary File 1 (Section 5)). Note that a  $p$ -value less than 0.05 was considered to represent the cutoff.

**Table 2.** General properties and date retrieved of the signaling networks. The number of differentially expressed genes (DEGs) is also given, which are those common between the edge list genes and gene expression profile genes and identified by the GEO/GDSC database as either up- or down-regulated. Samples are the total number of samples in GEO and GDSC databases for which expression data were available for the given gene pair. The node number of the giant component, the diameter of the network, and the ratio of shared genes between edge list genes and gene-expression-profile genes are presented in the last three columns.

	KEGG						OmniPath					
	Date Retrieved	DEGs	Samples	Giant Component	Diameter	Ratio	Date Retrieved	DEGs	Samples	Giant Component	Diameter	Ratio
<b>GEO</b>	2017.08	3047	40,903	2549	17	0.95	201,905	4724	40,774	3848	17	0.95
<b>GDSC</b>	2017.10	2745	1018	2583	17	0.16	201,905	4402	1018	4045	15	0.25





**Figure 2.** (A) An exemplary relationship between gene pair expression. These scatter plots contain the Pearson coefficient correlations and fitted linear regression line. The X-axis and Y-axis values differ according to the expression profile of this gene pair in different gene expression datasets. The gene expression profiles of these exemplary gene pairs in the edge list before pre-processing are depicted. The same gene pairs' expression profiles were separated according to the four relevant datasets. (B) The proportion of eligible and ineligible edges in the four parallel analyses. The numbers around each chart represent the number of edges at that point.

#### 2.4. Randomly Selected Unconnected Gene Pairs

The edge lists obtained in the previous step were converted into adjacency matrices using the *igraph* package in R [20]. Then, the adjacency matrix was self-multiplied more than the diameter of the network (e.g.,  $n > 17$ ) (Table 2). We then randomly selected 1000 unconnected gene pairs ten times, for which the corresponding elements in the matrix were zero (gene pairs with no direct immediate and non-immediate interactions). For these gene pairs, which we called unconnected gene pairs (UGPs), the same downstream analyses, i.e., pre-processing and correlation analysis, were implemented to compare the significance and sign of correlation coefficients to connected gene pairs (Supplementary File 1 (Section 6) and Supplementary Files 7).

#### 2.5. Complex Subgraphs

Since gene pairs are not isolated within the whole signaling networks, the larger subnetworks, which consisted of gene pair relationships, needed to be considered. Therefore, we extracted specific subgraphs from the signaling networks to investigate any association between gene expression profiles and the complex structures of relationships for each gene pair. DNFB (Dual Negative Feedback Loop), DPFBL1 (Dual Positive Feedback Loop 1), and DPFBL2 (Dual Positive Feedback Loop 2) are subgraphs of gene pairs that influence each other directly twice (see Table 1 for the full

names of subgraphs). These pairs are readily found by checking the source and target nodes in the edge lists (or upper and lower triangles in adjacency matrices). We then focused on the connected gene pairs, which also influence each other indirectly by a sequence of intermediate nodes. Following matrix self-multiplication, the weighted and unweighted adjacency matrices of the giant component of eligible edges in the signaling network were powered by the network radius magnitude. Considering that the network is directed and the adjacency matrix is not symmetric, the feed-forward and feedback loops, i.e., MNFBL1-2, MPFBL1-2, MFFL1-2, and MNFFL1-2, were determined (Table 1). For a more detailed explanation of procedures, see Supplementary File 1 (Sections 7 and 8).

### 3. Results

The overall details of the four parallel coherency analyses are presented, including the dimension of the expression matrices generated from whole-transcriptome expression profiles, and the size and diameter of the giant component in each analysis (Table 2). Notably, the number of DEGs was higher in OmniPath than KEGG, even though the size of the KEGG network is 1.8-fold larger than the OmniPath network. The ratio of eligible edges to all edges was calculated for all four analyses (see Figure 2B). The ratio of eligible edges in the OmniPath edge list also was higher than KEGG based on both GDSC and GEO databases. In addition, the ratio of eligible edges was higher in GDSC compared with GEO. This comparison may indicate the higher quality of gathered and annotated data in OmniPath and GDSC databases.

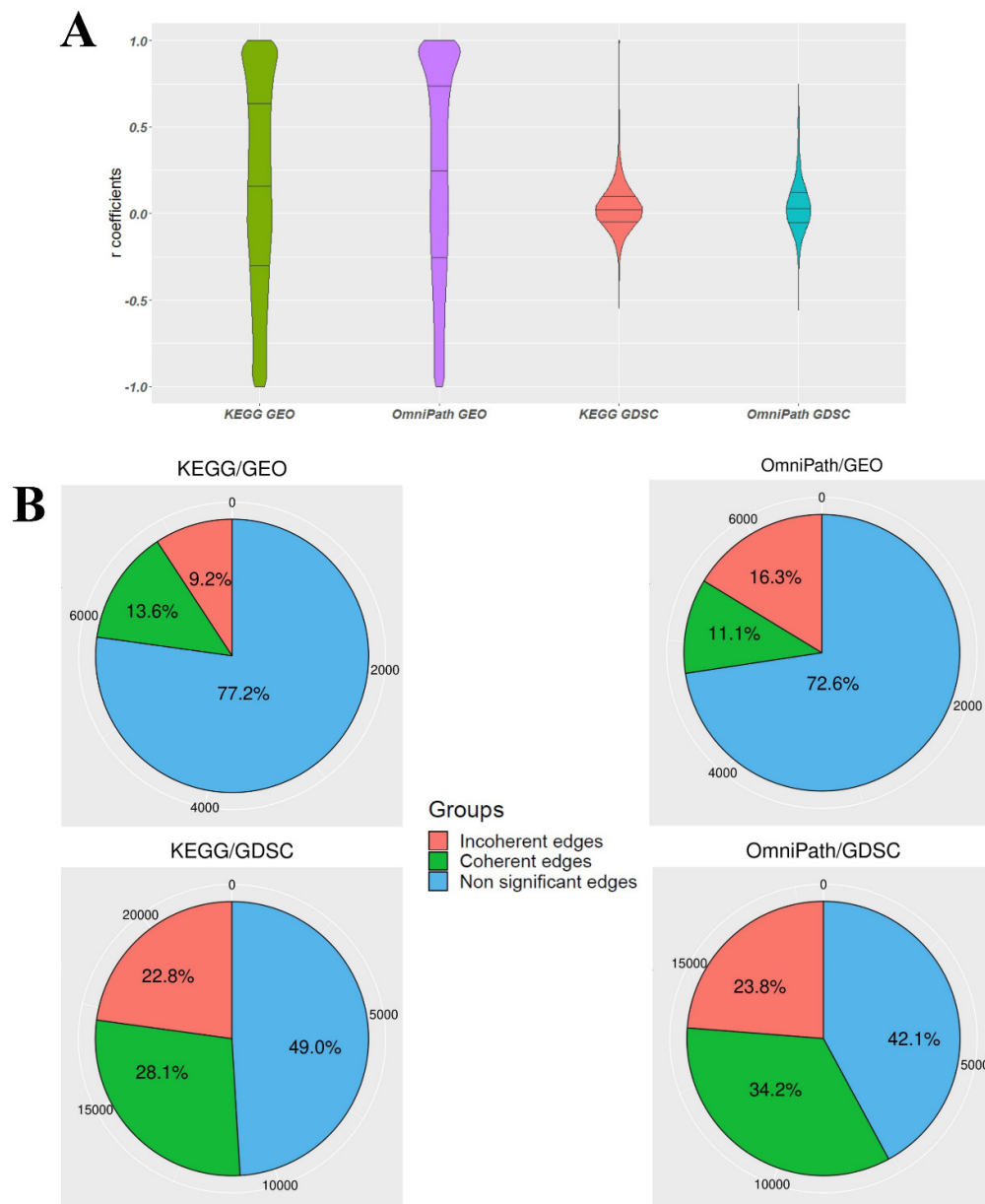
#### 3.1. The Ratio of Coherency for Gene Pairs

After filtering out heterogeneous edges, an extensive list of homogeneous edges was constructed (Supplementary File 1 (Sections 3.5–3.7) and Supplementary File 6) for correlation analysis. We performed correlation analysis in datasets with more than two samples, and the majority of the datasets had more than five samples. The violin plots of Pearson correlation coefficients for each analysis are shown in Figure 3A. The distribution of the coefficients showed a nearly uniform distribution with a little left skewness for KEGG/GEO and OmniPath/GEO, while for KEGG/GDSC and OmniPath/GDSC, it followed a normal distribution with the median at approximately zero. In addition to the issue of different sample sizes in GEO and GDSC, this suggests that for GDSC-based edges, correlations between the expression profiles of the gene pairs do not tend to show a high positive or negative correlation. In other words, for a given gene pair (I, II), over- or under-expression of gene I does not have a substantial effect on the expression of gene II regardless of the edge sign.

Figure 3B depicts the ratios of coherent and incoherent gene pairs along with the number of non-significant edges that have False Discovery Rate (FDR) adjusted *p*-values larger than 0.05, and we could not establish the coherency status at a likelihood greater than or equal to 95%. In addition, the sum of the ratio of incoherent gene pairs and non-significant edges was more than the ratio of coherent gene pairs in all four analyses. The ratio of coherent gene pairs in OmniPath was, in general, higher than for KEGG, while the ratio of coherent gene pairs in the GDSC database was higher than that of GEO. In more detail, when we considered GDSC as a cancer-specific homogenous dataset, the percent of coherent edges increased from 13.5% in GEO to 28.1% according to the KEGG database and from 11.1% in GEO to 34.2% based on the OmniPath database. In the following figures, the results correspond to the Pearson correlation coefficient. However, note that the ratio of coherency for gene pairs is also similar when using Spearman and Kendall rank correlation coefficients (Supplementary File 1 (Section 4)).

Figure 4 shows the FDR-adjusted *p*-values versus correlation coefficients of activation and inhibition edges in all four analyses. The symmetric pattern of coefficients is recognizable for both activation and inhibition edges in the four analyses. This suggests that the correlation between a given gene pair is not predominantly affected by the sign of the interaction. In other words, although activation edges illustrate an overrepresentation of strongly positively correlated gene pairs in all four analyses, the inhibition edges do not display any enrichment in the strong negative side of plots compared with the strong positive side. This also demonstrates that the majority of coherent gene pairs are activation rather than inhibition edges. In the next step, we tried to explore and provide

supporting evidence for the incoherent gene pairs by focusing on more complex subgraphs in the signaling network.

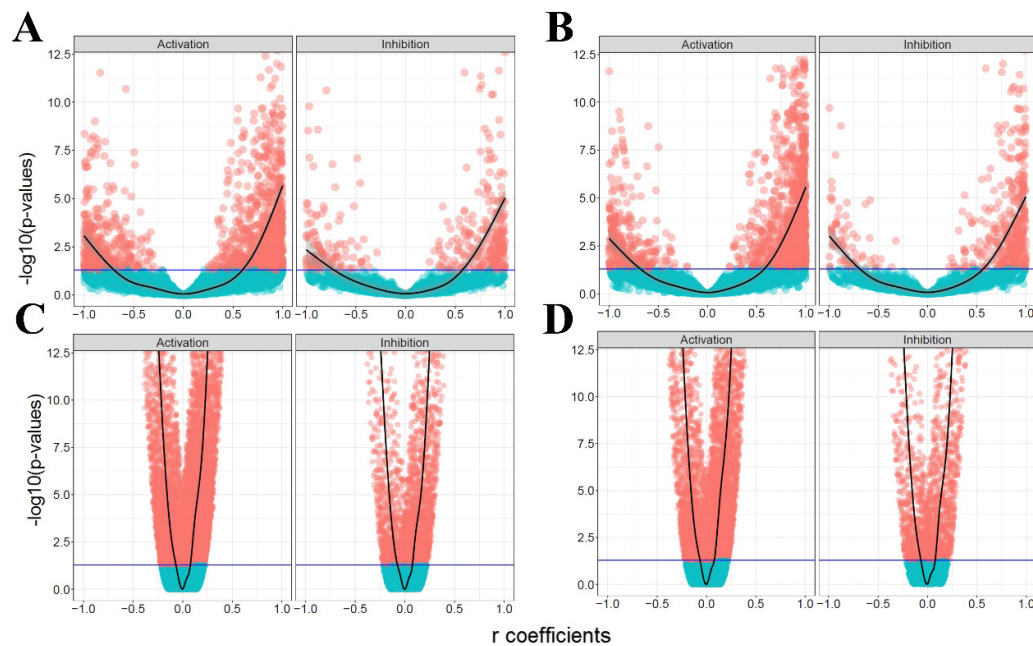


**Figure 3.** (A) Distribution of Pearson correlation coefficient values for the four parallel coherency analyses. (B) The ratio of coherent and incoherent edges in addition to the ratio of non-significant edges that have adjusted  $p$ -values of correlation test larger than 0.05. The values around each pie chart represent the exact numbers.

### 3.2. The Ratio of Coherency on Subgraphs

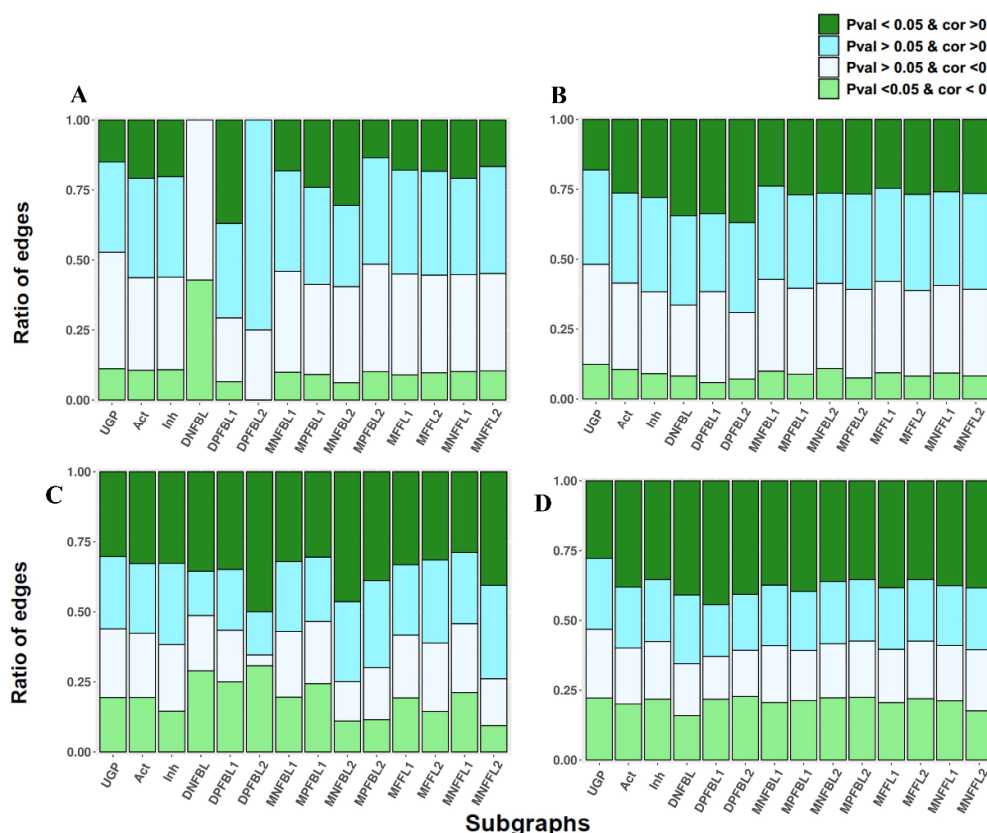
In this step, we explored whether complex subgraphs are coherent compared to when only single edges are considered. Briefly, we assumed that observing some incoherency of activation and inhibition edges may depend on the complex structure of the signaling network and, thus, the behavior of larger subgraphs should be considered to infer coherency (Figure 5). Like the simple activation or inhibition edges, correlations are computed and categorized by taking into account the correlation sign for each mentioned subgraph and the calculated FDR-adjusted  $p$ -values (see details

in Supplementary File 2 for all the four analyses). For example, we expected that the proportion of significant positive correlations is more than the negative correlations in DNFBFL as a negative feedback loop when compared with DPFBFL1 and DPFBFL2, because the two edges of the DNFBFL do not have the same sign, and overexpression of one protein entails the under-expression of the other one in a negative feedback loop. However, this expectation was only partially fulfilled in KEGG/GEO analysis.



**Figure 4.** The volcano plots of the activation and inhibition edges. The horizontal axis is the Pearson correlation coefficient, and the vertical axis shows log transformed False Discovery Rate (FDR)-adjusted  $p$ -values. The threshold line (blue) represents the significance cut-off value of 0.05. (A–D) plots correspond to KEGG/GEO, OmniPath/GEO, KEGG/GDSC, and OmniPath/GDSC analyses, respectively.

To statistically compare the coherency of different subgraphs, the correlation analysis was also implemented on multiple sets of 1000 randomly unconnected gene pairs (UGPs), and a binomial proportion test was then undertaken to compare all of the proportions as illustrated in Figure 5. UGP sets were employed to compare the results of the coherency analysis with the ones obtained from gene pairs where relationships exist between them. In most cases (>60%), the comparison showed a statistically significant difference between pairwise proportions of UGPs and the connected gene pairs, e.g., Act, Inh, DNFBFL for all four combinations of databases (Supplementary File 3). It is possible that the connected genes are affected by each other with respect to UGPs, but it may happen in a more complex way that it is not inferred by correlation analysis (Figure 5). We also aimed to continue our search to check coherency in larger subgraph structures. We therefore identified subgraphs that contained more than two edges, i.e., MNFBFLs (Multiple Negative Feedback Loop 1 and 2), MPFBFLs (Multiple Positive Feedback Loop 1 and 2), MFFFLs (Multiple Feed-Forward Loop 1 and 2), and MNFFFLs (Multiple Negative Feed-Forward Loop 1 and 2), using matrix self-multiplication (See Table 1). The purpose of self-multiplication was to retrieve the related gene pairs in complex subgraphs and unrelated gene pairs in unconnected gene pairs. However, our results demonstrated the limitation of strong coherent relationships among gene pairs, suggesting that, independent of the structure of the subgraph, gene expression profiles are discordant with the functionality of signaling circuits.



**Figure 5.** The ratio of eligible and homogeneous edges involved in different subgraphs are represented by stacked bar plots for all four analyses. (A,B) are KEGG/GEO and OmniPath/GEO plots, and (C,D) plots correspond to KEGG/GDSC and OmniPath/GDSC. Note that the unconnected gene pair (UGP) ratios are the average ratio over of all UGP sets. See Table 1 for the full names of subgraphs.

#### 4. Discussion

Incoherency in signaling networks at the transcript level might be a strategy used by cells to tune both genetic and environmental signals or a byproduct resulting from frequent modulation of signal transduction within biological systems. In fact, cells acquire the characteristics they need by setting different layers of gene regulations, which illustrates the importance of different layers of post transcriptional regulation [21]. High-throughput technologies, such as mRNA microarray, CHIP-seq (Chromatin Immunoprecipitation Sequencing), and mass spectrometry proteomics, have uncovered the amount of expression through determination of mRNA and protein levels [22]. There is an apparent correspondence between mRNA and protein concentrations [23]. Nonetheless, more than fifty percent of protein variation cannot be explained by variation in mRNA concentration [23–25]. These unexplained variations might come from organism-specific translational and post-translational regulations, including protein degradation by ubiquitylation and sumoylation [26]. Because we were dealing with whole-human signaling networks, considering all aspects that impact the expression of each gene is far-fetched. Such an analysis would represent the total coherency based on the analysis of gene expression at the mRNA level. The correlation between transcript and protein concentrations are considerable for some house-keeping genes. However, in many eukaryotes, there is no strong correlation for genes of signal transduction or transcriptional regulation, while their encoded proteins are often involved in different signaling networks and determine cell fate and behavior of the system [27].

Although the regulation of gene expression results in a particular concentration of proteins, it is not sufficient to completely describe protein abundances [28,29]. The roles of other mechanisms, such as post-transcriptional, post-translational, and protein degradation regulations, have been reported

to control steady-state protein abundance and protein activity [28,30]. For example, miRNAs are associated with proteins and diseases as regulators of gene expression [31,32]. There are several studies focusing on the regulatory role of long noncoding RNAs (lncRNAs) in wound healing and cancers [33–35]. Alternative splicing also results in diverse forms of proteins, and the functions of proteins and corresponding signal transduction pathways are affected by this type of regulation in some diseases [36]. Moreover, numerous methylation sites have been discovered on human proteins that impact cell signaling [37]. Therefore, the incoherency might come from the effects of complicated regulatory networks to buffer signaling networks within biological systems. These modifications are thus likely to be the source of poor coherency between signal transduction and gene expression and are particularly valid for multicellular eukaryotes such as worm and fly. In contrast, yeast genes engaged in signal transduction exhibit high correlation between mRNA and protein concentrations [24]. Interestingly, Larsen et al. recently demonstrated that there is no causal relationship between the expression of transcription factors and their targets in the gene regulatory network of *Escherichia coli* and, therefore, the transcriptional regulation cannot be adequately addressed by examining the current static gene regulatory networks [38]. As a result, inferring gene regulatory or signaling relationships from transcript data is challenging because these data are not the proxy of molecular activity. Only in some cases are the results acceptable for constructing logical circuits of biological elements, e.g., if the components of the system are all kinases and the transition of the signals is related to the phosphorylation process [10,39,40]. Different layers of gene regulation need to be considered for each gene pair when inferring causal relationships between a pair of genes. Previously, we also demonstrated the association of altered expression with the signaling circuits in chronic obstructive pulmonary disease [41] and rabies infection [42] as case studies.

We implemented three different types of correlation analyses, namely Pearson, Spearman, and Kendall, all of which exhibited similar findings. Based on the correlation results in Figure 3A, the volcano plots in Figure 4, which exhibit no significant difference between activation and inhibition edges, and the ratios in Figure 5, causal relationship can be inferred poorly at the transcript level, at least in multicellular eukaryotic organisms such as *Homo sapiens*. Proportional binomial tests suggested that there are statistically significant differences between UGPs and other subgraphs (Supplementary File 3), and this demonstrates that the structure of subgraphs affects the coherency. Based on these considerations, it is therefore strongly advocated that information in signaling networks be used or that relationships between the genes are defined in addition to assessing gene expression at both transcript and protein levels.

## 5. Conclusions

Although there is a general assumption that the expression level could strengthen or weaken the signal to transduce in the signaling pathways, we show that the mRNA level of gene pairs is generally incoherent with the way they manipulate one another (Figure 5). We also showed that there is a level of association between the structure of the subgraphs and gene pair expression profiles. Expression profiles of the unconnected gene pairs were statistically more independent than connected ones. Note that these observations are according to the four datasets used here, and independent investigation of gene pair relationships based on more accurate and specific datasets is recommended.

In this study, we aimed to focus on the impact of the type of relationship on any given stimulated signaling pathway, an area which is usually ignored in functional genomic studies. We demonstrate that DEGs have only partial information on the whole story of the associated mechanism. The presented findings support the idea of using interaction- or edge-centric (based on DEG–DEG activatory or inhibitory relationships) enrichment analysis against node-centric (based only on a list of DEGs) analyses to provide a more vivid perspective of implicated pathways. This is because the majority of the altered expression of DEGs gradually disappears and is then overlooked by the whole signaling network system, whether stimulated endogenously or exogenously.

**Supplementary Materials:** The following are available online at <https://zenodo.org/record/3822611#.Xr0F52gzbIU>, Supplementary File 1: The experimental procedure based on KEGG/GEO analysis in detail. This file contains nine sections. The first section describes how the KEGG edge

list with 26,490 edges was built. Next, in the second section, downloading and merging the up-down gene expression profiles is explained for KEGG genes. Section three walks you through the preprocessing of the expression profiles. In this step, an extensive list containing 1969 experiments (GDS) was built. A large expression matrix called Exprtable with 40,903 samples in column and 3187 genes in rows was constructed. From this matrix, a list called SignalingNet was constructed, with an element for each gene pair in the KEGG edge list. In the fourth section, each element of SignalingNet contains the expression values and correlation information for the source and the target genes. Section five includes the information for coherency of the edges and the number of activation and inhibition edges having a specific  $p$ -value and correlation coefficient. Then, in the sixth section, ten sets of 1000 unconnected node pairs were built in which the genes never reach one another (based on KEGG information). The correlation analysis was also performed on these node pairs. In the seventh section, the number of edges having a specific  $p$ -value and correlation coefficient engaged in two-edge subgraphs was computed. Afterward, in the eighth analysis, the number of edges having a specific  $p$ -value and correlation coefficient engaged in multiple-edge subgraphs were computed. Finally, in the ninth section, the results are summarized in some tables. Supplementary File 2: Correlation analysis of all four analyses. Results are the number of edges having specific  $p$ -values and correlations in different subgraphs. Supplementary File 3: The proportional statistical tests between the rows in the tables in supplementary File 2 for all four analysis in separate sheets. Supplementary File 4: The KEGG and OmniPath edge lists. Supplementary File 5: The large expression matrices constructed based on four analysis analyses. Supplementary File 6: The SignalingNet list for the four analyses. Supplementary File 7: The unconnected SignalingNet list for the four analyses.

**Author Contributions:** M.M., N.A.P., J.T., and M.J. contributed to conceptualization and investigation that it means they conceived and commenced the project and provided direction and feedback on the final results. M.P. contributed to data curation and formal analysis, and he wrote the script codes and gathered datasets. MP1 and Mehrdad P. prepared the original draft of this study. R.K. and J.A. carried out some part of data curation, acquisition, and computational analysis. N.A.P., J.T. and M.J. revised and edited the manuscript critically. All authors read and approved the final manuscript.

**Funding:** This research was supported by Research Core: “Bio-Mathematics with computational approach” of Tarbiat Modares University with grant number “IG-39706”, the Academy of Finland with grant number 317680 and European Research Council with grant number 716063.

**Acknowledgments:** The authors would like to thank Julio Saez-Rodriguez, Bence Szalai, and Zahra Razaghi-Moghadam for their valuable comments and discussions.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Liu, A.; Trairatphisan, P.; Gjerga, E.; Didangelos, A.; Barratt, J.; Saez-Rodriguez, J. From expression footprints to causal pathways: Contextualizing large signaling networks with CARNIVAL. *NPJ Syst. Biol. Appl.* **2019**, *5*, 40.
2. Ma’ayan, A. Introduction to network analysis in systems biology. *Sci. Signal.* **2011**, *4*, tr5.
3. Hecker, M.; Lambeck, S.; Toepfer, S.; Van Someren, E.; Guthke, R. Gene regulatory network inference: Data integration in dynamic models—A review. *Biosystems* **2009**, *96*, 86–103.
4. Shipley, B. *Cause and Correlation in Biology: A User’s Guide to Path Analysis, Structural Equations and Causal Inference with R*; Cambridge University Press: Cambridge, UK, 2016.
5. Bansal, M.; Belcastro, V.; Ambesi-Impombato, A.; Di Bernardo, D. How to infer gene networks from expression profiles. *Mol. Syst. Biol.* **2007**, *3*, 78.
6. Margolin, A.A.; Nemenman, I.; Basso, K.; Wiggins, C.; Stolovitzky, G.; Favera, R.D. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinform.* **2006**, *7*, S7.
7. Marbach, D.; Costello, J.C.; Küffner, R.; Vega, N.M.; Prill, R.J.; Camacho, D.M. Wisdom of crowds for robust gene network inference. *Nat. Methods* **2012**, *9*, 796–804.
8. Lin, Z.; Guo, H.; Cao, Y.; Zohrabian, S.; Zhou, P.; Ma, Q. Acetylation of VGLL4 regulates Hippo-YAP signaling and postnatal cardiac growth. *Dev. Cell* **2016**, *39*, 466–479.
9. Song, Y.; Li, Z.-X.; Liu, X.; Wang, R.; Li, L.-W.; Zhang, Q. The Wnt/ $\beta$ -catenin and PI3K/Akt signaling pathways promote EMT in gastric cancer by epigenetic regulation via H3 lysine 27 acetylation. *Tumor Biol.* **2017**, *39*, 1010428317712617.



10. Hill, S.M.; Heiser, L.M.; Cokelaer, T.; Unger, M.; Nesser, N.K.; Carlin, D.E. Inferring causal molecular networks: Empirical assessment through a community-based effort. *Nat. Methods* **2016**, *13*, 310–318.
11. Jafari, M.; Ansari-Pour, N.; Azimzadeh, S.; Mirzaie, M. A logic-based dynamic modeling approach to explicate the evolution of the central dogma of molecular biology. *PLoS ONE* **2017**, *12*, e0189922.
12. Costa-Silva, J.; Domingues, D.; Lopes, F.M. RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS ONE* **2017**, *12*, e0190152.
13. Han, J.; Shi, X.; Zhang, Y.; Xu, Y.; Jiang, Y.; Zhang, C. ESEA: Discovering the dysregulated pathways based on edge set enrichment analysis. *Sci. Rep.* **2015**, *5*, 13044.
14. Edgar, R.; Domrachev, M.; Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **2002**, *30*, 207–210.
15. Yang, W.; Soares, J.; Greninger, P.; Edelman, E.J.; Lightfoot, H.; Forbes, S. Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **2012**, *41*, D955–D961.
16. Kanehisa, M.; Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30.
17. Túrei, D.; Korcsmáros, T.; Saez-Rodriguez, J. OmniPath: Guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **2016**, *13*, 966.
18. Zhang, J.D.; Wiemann, S. KEGGgraph: A graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics* **2009**, *25*, 1470–1471.
19. Luke, D.A. *A User's Guide to Network Analysis in R*; Springer: Cham, Switzerland, 2015.
20. Csardi, G.; Nepusz, T.J.I. The igraph software package for complex network research. *InterJ. Complex Syst.* **2006**, *1695*, 1–9.
21. Koussounadis, A.; Langdon, S.P.; Um, I.H.; Harrison, D.J.; Smith, V.A. Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. *Sci. Rep.* **2015**, *5*, 10775.
22. Vogel, C.; Marcotte, E.M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **2012**, *13*, 227–232.
23. Liu, Y.; Beyer, A.; Aebersold, R. On the dependency of cellular protein levels on mRNA abundance. *Cell* **2016**, *165*, 535–550.
24. de Sousa Abreu, R.; Penalva, L.O.; Marcotte, E.M.; Vogel, C. Global signatures of protein and mRNA expression levels. *Mol. Biosyst.* **2009**, *5*, 1512–1526.
25. Vogel, C.; de Sousa Abreu, R.; Ko, D.; Le, S.Y.; Shapiro, B.A.; Burns, S.C. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.* **2010**, *6*, 400.
26. Lu, P.; Vogel, C.; Wang, R.; Yao, X.; Marcotte, E.M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* **2007**, *25*, 117–124.
27. MacKay, V.L.; Li, X.; Flory, M.R.; Turcott, E.; Law, G.L.; Serikawa, K.A. Gene expression analyzed by high-resolution state array analysis and quantitative proteomics response of yeast to mating pheromone. *Mol. Cell. Proteom.* **2004**, *3*, 478–489.
28. Mata, J.; Marguerat, S.; Bähler, J. Post-transcriptional control of gene expression: A genome-wide perspective. *Trends Biochem. Sci.* **2005**, *30*, 506–514.
29. Jafari, M.; Mirzaie, M.; Sadeghi, M.; Marashi, S.-A.; Rezaei-Tavirani, M. Exploring biological processes involved in embryonic stem cell differentiation by analyzing proteomic data. *Biochim. Biophys. Acta (BBA)-Proteins Proteom.* **2013**, *1834*, 1063–1069.
30. Ansari-Pour, N.; Razaghi-Moghadam, Z.; Barneh, F.; Jafari, M. Testis-Specific Y-Centric Protein-Protein Interaction Network Provides Clues to the Etiology of Severe Spermatogenic Failure. *J. Proteome Res.* **2016**, *15*, 1011–1022.
31. Inui, M.; Martello, G.; Piccolo, S. MicroRNA control of signal transduction. *Nat. Rev. Mol. Cell Biol.* **2010**, *11*, 252–263.
32. Mørk, S.; Pletscher-Frankild, S.; Pallega, A.; Gorodkin, J.; Jensen, L.J. Protein-driven inference of miRNA-disease associations. *Bioinformatics* **2014**, *30*, 392–397.
33. Wang, J.; Song, Y.-X.; Ma, B.; Wang, J.-J.; Sun, J.-X.; Chen, X.-W. Regulatory roles of non-coding RNAs in colorectal cancer. *Int. J. Mol. Sci.* **2015**, *16*, 19886–19919.
34. Gibb, E.A.; Brown, C.J.; Lam, W.L. The functional role of long non-coding RNA in human carcinomas. *Mol. Cancer* **2011**, *10*, 38.



35. Yu, B.; Zhou, S.; Yi, S.; Gu, X. The regulatory roles of non-coding RNAs in nerve injury and regeneration. *Prog. Neurobiol.* **2015**, *134*, 122–139.
36. Kelemen, O.; Convertini, P.; Zhang, Z.; Wen, Y.; Shen, M.; Falaleeva, M. Function of alternative splicing. *Gene* **2013**, *514*, 1–30.
37. Biggar, K.K.; Li, S.S.-C. Non-histone protein methylation as a regulator of cellular signalling and function. *Nat. Rev. Mol. Cell Biol.* **2015**, *16*, 5–17.
38. Larsen, S.J.; Röttger, R.; Schmidt, H.H.H.W.; Baumbach, J. E. coli gene regulatory networks are inconsistent with gene expression data. *Nucleic Acids Res.* **2019**, *47*, 85–92.
39. Tang, J.; Gautam, P.; Gupta, A.; He, L.; Timonen, S.; Akimov, Y. Network pharmacology modeling identifies synergistic Aurora B and ZAK interaction in triple-negative breast cancer. *NPJ Syst. Biol. Appl.* **2019**, *5*, 20.
40. D'Souza, R.C.; Knittle, A.M.; Nagaraj, N.; van Dinther, M.; Choudhary, C.; Ten Dijke, P. Time-resolved dissection of early phosphoproteome and ensuing proteome changes in response to TGF- $\beta$ . *Sci. Signal.* **2014**, *7*, rs5.
41. Azimzadeh, S.; Mirzaie, M.; Jafari, M.; Mehrani, H.; Shariati, P.; Khodabandeh, M. Signaling network of lipids as a comprehensive scaffold for omics data integration in sputum of COPD patients. *Biochim. Biophys. Acta (BBA)-Mol. Cell Biol. Lipids* **2015**, *1851*, 1383–1393.
42. Azimzadeh Jamalkandi, S.; Mozhgani, S.-H.; Gholami Pourbadie, H.; Mirzaie, M.; Noorbakhsh, F.; Vaziri, B. Systems Biomedicine of Rabies Delineates the Affected Signaling Pathways. *Front. Microbiol.* **2016**, *7*, 1688.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).